Statistica Sinica **24** (2014), 723-747
doi:http://dx.doi.org/10.5705/ss.2012.254

EMPIRICAL LIKELIHOOD FOR ESTIMATING EQUATIONS WITH NONIGNORABLY MISSING DATA

Niansheng Tang[1], Puying Zhao[1] and Hongtu Zhu[2]

[1] *Yunnan University and* [2] *University of North Carolina at Chapel Hill.*

Abstract: We develop an empirical likelihood (EL) inference on parameters in generalized estimating equations with nonignorably missing data. We consider an exponential tilting model for the nonignorably missing mechanism, and propose modified estimating equations by imputing missing data through a kernel regression method. We establish some asymptotic properties of the maximum EL estimators of the unknown parameters under different scenarios. With the use of auxiliary information, the maximum EL estimators are statistically more efficient. Simulation studies are used to assess the finite sample performance of our proposed maximum EL estimators. We apply the proposed maximum EL estimators to investigate a data set on earnings obtained from the New York Social Indicators Survey.

Key words and phrases: Empirical likelihood, estimating equations, exponential tilting, imputation, kernel regression, nonignorable missing data.

1. Introduction

Missing data are encountered in various settings, including surveys, clinical trials and longitudinal studies (Little and Rubin (2002)); responses and/or covariates may be missing in practice. Statistical models for dealing with the missing data depend on a missing data mechanism such as data not missing at random (NMAR), also referred to as nonignorable missingness. For example, when there are NMAR responses, the complete-case analysis can result in biased and inefficient parameter estimates, whereas to incorporate additional information from incomplete cases, one needs to assume a parametric (or semiparametric) model for the missing data mechanism. However, the assumptions underlying such NMAR models are difficult to verify in practice and the resulting estimates and tests may be sensitive to these assumptions. This paper develops an inference procedure for parameters in generalized estimating equations (GEEs) with nonignorably missing data.

Various generalized estimating equations have been developed to handle missing data, that are either missing at random (MAR) or NMAR, primarily due to their being robust against model misspecification. Robins, Rotnitzky, and Zhao (1994) developed a class of estimators based on inverse probability-weighted

estimating equations (EEs) when the probability of missingness is modeled parametrically, and Robins, Rotnitzky, and Zhao (1995) proved the semiparametric efficiency bound for parameter estimation. Lipsitz, Ibrahim, and Zhao (1999) presented an EM algorithm to estimate parameters defined by a weighted EE with missing covariate data. It is more challenging to deal with NMAR data due to the unverifiable assumptions introduced by the statistical models for it. Troxel, Lipsitz, and Brennan (1997) proposed weighted EEs for data with nonignorable nonresponse to eliminate the biases in the complete-case analysis that ignores missing cases when the weights (the inverse probability of being observed) are estimable. Kim and Yu (2011) developed an exponential tilting model and proposed a semiparametric estimation method of mean functionals with nonignorably missing data. See Ibrahim et al. (2005) and Ibrahim and Molenberghs (2009) for a detailed overview and comparisons of various paradigms for handling missing data. All these methods are developed on the basis of non-empirical likelihood.

There is considerable interest in the development of EL for GEEs with/without ignorably missing data. Empirical likelihood allows one to employ likelihood methods in a nonparametric or semiparametric setting. It has been shown that EL has various advantages over other competing methods, including generalized method of moments (GMM) (Newey and Smith (2004)). Compared with EE, EL allows the easy incorporation of auxiliary information and the number of estimating equations can be greater than the number of parameters. See, for example, Qin and Lawless (1994); Zhou, Wan, and Wang (2008); Zhu et al. (2008); Wang and Chen (2009), and Qin, Zhang, and Leung (2009), among many others. Zhou, Wan, and Wang (2008) proposed a kernel-assisted EE imputation scheme and used EL and GMM on parameters in GEEs. Wang and Chen (2009) proposed a nonparametric imputation method to remove the selection bias in the missingness and showed that the maximum EL estimators can be efficient. However, little has been done on the development of the EL method for GEEs with nonignorably missing data.

We develop a general EL inference procedure for parameters in the GEEs with nonignorably missing data. We integrate the modeling of nonignorably missing data, the EL method, and the imputation of EEs by using the observed data rather than imputing the missing data. Specifically, we consider the exponential tilting model with known and estimated tilting parameters as the missing mechanism for nonignorably missing data, which leads to a more robust estimator. We extend the estimation of mean functionals with nonignorably missing data to the estimation of parameters in GEEs. We systematically investigate the asymptotic properties of the maximum EL estimators under this new setting.

The rest of this article is organized as follows. In Section 2, we describe the proposed kernel-assisted EE imputation scheme based on the exponential tilting

model of nonignorably missing data. As well, we outline the formulations of EL with and without auxiliary information by utilizing the imputation scheme. In Section 3, we establish the asymptotic properties of the proposed maximum EL estimators. Two simulation studies and a data analysis are used to compare the finite sample performance of the proposed maximum EL estimators with competing methods, in Section 4. Technical details are given in the Appendix.

2. Methods

2.1. Imputation based on the exponential tilting model

Let $\{U_i = (Z_i^T, Y_i^T)^T : i = 1, \ldots, n\}$ be a set of independent and identically distributed random vectors from a distribution $F(z, y)$, where the Z_i's are d_x-dimensional and observable, and the Y_i's are d_y-dimensional and subject to missingness. Generally, the missing components may vary across different individuals. For simplicity, we assume that the missing components have the same components for $U_1, \ldots, U_n$. Furthermore, a missing variable Y_i may represent a response or covariate. Without assuming a specific form for $F(u)$, we are interested in making statistical inference on a $p \times 1$ vector, denoted by θ, based on $q(\geq p)$ functionally independent EEs $\psi(Y_i, Z_i; \theta) = (\psi_1(Y_i, Z_i; \theta), \ldots, \psi_q(Y_i, Z_i; \theta))^T$ that satisfy the unconditional moment condition of the form $E_F\{\psi(Y_i, Z_i; \theta_0)\} = 0$ for $\theta_0 \in \Theta \subset \mathcal{R}^p$, where θ_0 is the true value and E_F denotes the expectation with respect to F. The Y_i's are assumed to be nonignorably missing. Let X_i be Z_i or a subset of Z_i, and let $\delta_i = 1$ if Y_i is observed and $\delta_i = 0$ if Y_i is missing. It is assumed that δ_i and δ_j are independent for any $i \neq j$ and δ_i depends on X_i and Y_i such that $P(\delta_i = 1|X_i, Y_i) \stackrel{\Delta}{=} \pi(X_i, Y_i)$ for $i = 1, \ldots, n$. When $\pi(X_i, Y_i)$ depends on the value of Y_i, it is the NMAR condition of Little and Rubin (2002).

We consider an exponential tilting model for nonignorably missing data Y_i's given by

$$\text{logit}\{\pi(X_i, Y_i) \stackrel{\Delta}{=} P(\delta_i = 1|X_i, Y_i)\} = g(X_i) + \phi Y_i \tag{2.1}$$

for some unknown function $g(\cdot)$ and ϕ, where logit denotes the logit function. When $\phi = 0$, (2.1) becomes an MAR model. Let $f_1(Y_i|X_i)$ be the conditional density of Y_i given X_i and $\delta_i = 1$, and let $f_0(Y_i|X_i)$ be the conditional density of Y_i given X_i and $\delta_i = 0$. Then, by following the reasoning of Kim and Yu (2011), we have

$$f_0(Y_i|X_i) = f_1(Y_i|X_i) \times \frac{\exp(\gamma Y_i)}{E\{\exp(\gamma Y_i)|X_i, \delta_i = 1\}}, \tag{2.2}$$

where $\gamma = -\phi$ is an unknown tilting parameter that measures the amount of departure from the MAR assumption. When $\gamma = 0$, (2.2) reduces to $f_0(Y_i|X_i) = f_1(Y_i|X_i)$.

To incorporate the incomplete cases, we consider a new set of EEs given by

$$\tilde{\psi}(Y_i, Z_i; \theta) = \delta_i \psi(Y_i, Z_i; \theta) + (1 - \delta_i) m_\psi(X_i; \theta), \tag{2.3}$$

where $m_\psi(X_i; \theta) = E_F\{\psi(Y_i, Z_i; \theta)|X_i\}$. Under MAR, the EEs in (2.3) reduce to the estimating equations of Zhou, Wan, and Wang (2008). Since $m_\psi(X_i; \theta)$ defined in (2.3) is unknown, it is necessary to estimate (or 'impute') $m_\psi(X_i; \theta)$ from the observed data set. Under the MAR assumption, a consistent estimator of $m_\psi(X_i; \theta)$ can be obtained from a consistent estimator of $m_{1\psi}(X_i; \theta) = E\{\psi(Y_i, Z_i; \theta)|X_i, \delta_i = 1\}$, denoted by $\hat{m}_{1\psi}(X_i; \theta)$. Substituting $\hat{m}_{1\psi}(X_i; \theta)$ in (2.3) leads to $\hat{\psi}_1(Y_i, Z_i; \theta) = \delta_i \psi(Y_i, Z_i; \theta) + (1 - \delta_i)\hat{m}_{1\psi}(X_i; \theta)$, which is biased under NMAR.

We construct a consistent estimator of $m^0_\psi(X_i; \theta) = E\{\psi(Y_i, Z_i; \theta)|X_i, \delta_i = 0\}$. Under the NMAR assumption, it is difficult to estimate $m^0_\psi(X_i; \theta)$ due to the presence of missing data. It follows from (2.2) that

$$m^0_\psi(X_i; \theta) = \frac{E\{\delta_i \psi(Y_i, Z_i; \theta) \exp(\gamma Y_i)|X_i\}}{E\{\delta_i \exp(\gamma Y_i)|X_i\}}. \tag{2.4}$$

Then, under the NMAR assumption, we construct a set of EEs for $\psi(Y_i, Z_i; \theta)$ given by

$$\hat{\psi}(Y_i, Z_i; \theta) = \delta_i \psi(Y_i, Z_i; \theta) + (1 - \delta_i) m^0_\psi(X_i; \theta), \tag{2.5}$$

where $m^0_\psi(X_i; \theta)$ is defined in (2.4) based on a tilting parameter γ.

If the exponential tilting model (2.1) is true, then we have

$$\begin{aligned}
&E\{\hat{\psi}(Y_i, Z_i; \theta)\} \\
&\quad = E\{\delta_i \psi(Y_i, Z_i; \theta) + (1 - \delta_i) m^0_\psi(X_i; \theta)\} \\
&\quad = E\Bigg\{ \mathrm{pr}(\delta_i = 1|X_i) E(\psi(Y_i, Z_i; \theta)|\delta_i = 1, X_i) \\
&\qquad + \mathrm{pr}(\delta_i = 0|X_i) \frac{E\{\delta_i \psi(Y_i, Z_i; \theta) \exp(\gamma Y_i)|X_i\}}{E\{\delta_i \exp(\gamma Y_i)|X_i\}} \Bigg\} \\
&\quad = E\Bigg\{ \mathrm{pr}(\delta_i = 1|X_i) E(\psi(Y_i, Z_i; \theta)|\delta_i = 1, X_i) \\
&\qquad + \mathrm{pr}(\delta_i = 0|X_i) E(\psi(Y_i, Z_i; \theta)|\delta_i = 0, X_i) \Bigg\} \\
&\quad = E\Bigg\{ E(\delta_i \psi(Y_i, Z_i; \theta)|X_i) + \mathrm{pr}(\delta_i = 0|X_i) \frac{E\{(1 - \delta_i)\psi(Y_i, Z_i; \theta)|X_i\}}{E\{(1 - \delta_i)|X_i\}} \Bigg\} \\
&\quad = E\{\psi(Y_i, Z_i; \theta)\} = 0.
\end{aligned}$$

The second equality holds since

$$\begin{aligned}\frac{E\{\delta\psi(Y,Z;\theta)\exp(\gamma Y)|X\}}{E\{\delta\exp(\gamma y)|X\}} &= \frac{E\{\pi(X,Y)\psi(Y,Z;\theta)\exp(\gamma Y)|X\}}{E\{\pi(X,Y)\exp(\gamma Y)|X\}}\\ &= \frac{E\{\psi(Y,Z;\theta)(1+\exp(g(X)-\gamma Y))^{-1}|X\}}{E\{(1+\exp(g(X)-\gamma Y))^{-1}|X\}}\\ &= \frac{E\{(1-\delta)\psi(Y,Z;\theta)|X\}}{E\{(1-\delta)|X\}} = E(\psi(Y,Z;\theta)|X,\delta=0).\end{aligned}$$

Thus (2.5) is unbiased, which is the key idea of our approach. From (2.1), we have $\pi(X_i,Y_i)=\{1+\exp(-g(X_i))\exp(\gamma Y_i)\}^{-1}$ with $\gamma=-\phi$ and $E\{\delta\exp(\gamma Y_i)|X_i\}=\exp(g(X_i))E(1-\delta_i|X_i)$, which indicates that

$$\exp(-g(X_i))=\frac{E(1-\delta_i|X_i)}{E\{\delta\exp(\gamma Y_i)|X_i\}}=\frac{\mathrm{pr}(\delta_i=0|X_i)}{\mathrm{pr}(\delta_i=1|X_i)E\{\exp(\gamma Y_i)|X_i,\delta_i=1\}}.$$

Then, we also have

$$\begin{aligned}E\{\psi(Y_i,Z_i;\theta)\} &= E\Big\{\frac{\delta_i\psi(Y_i,Z_i;\theta)}{\pi(X_i,Y_i)}\Big\}\\ &= E\{\delta_i\psi(Y_i,Z_i;\theta)+(1-\delta_i)m_\psi^0(X_i;\theta)\}=0.\end{aligned}$$

The equality holds since

$$\begin{aligned}E\Big\{\frac{\delta_i\psi(Y_i,Z_i;\theta)}{\pi(X_i,Y_i)}\Big\} &= E\Big\{\delta_i\psi(Y_i,Z_i;\theta)\Big[1+\frac{\mathrm{pr}(\delta_i=0|X_i)\exp(\gamma Y_i)}{\mathrm{pr}(\delta_i=1|X_i)E\{\exp(\gamma Y_i)|X_i,\delta_i=1\}}\Big]\Big\}\\ &= E\Big\{\mathrm{pr}(\delta_i=1|X_i)E(\psi(Y_i,Z_i;\theta)|\delta_i=1,X_i)\\ &\quad +\mathrm{pr}(\delta_i=0|X_i)\frac{E\{\psi(Y_i,Z_i;\theta)\exp(\gamma Y_i)|\delta_i=1,X_i\}}{E\{\exp(\gamma Y_i)|\delta_i=1,X_i\}}\Big\}\\ &= E\{\delta_i\psi(Y_i,Z_i;\theta)+(1-\delta_i)m_\psi^0(X_i;\theta)\}\\ &= E\{\psi(Y_i,Z_i;\theta)\}=0.\end{aligned}$$

This equality also holds under the MAR assumption.

Let $K(\cdot)$ be a d_x-dimensional kernel function of the m-th order satisfying $\int K(u_1,\ldots,u_{d_x})du_1\ldots du_{d_x}=1$, $\int u_s^l K(u_1,\ldots,u_{d_x})du_1\ldots du_{d_x}=0$ for any $s=1,\ldots,d_x$ and $1\le l<m$, and $\int u_s^m K(u_1,\ldots,u_{d_x})du_1\ldots du_{d_x}\neq 0$. Then, a nonparametric regression estimator of $m_\psi^0(X;\theta)=E\{\psi(Y,Z;\theta)|X,\delta=0\}$ can be written as

$$\hat m_\psi(X;\theta,\gamma)=\sum_{i=1}^n\omega_{i0}(X;\gamma)\psi(Y_i,Z_i;\theta),\tag{2.6}$$

where $\omega_{i0}(X;\gamma)=\delta_i\exp(\gamma Y_i)K_h(X-X_i)/\{\sum_{k=1}^n\delta_k\exp(\gamma Y_k)K_h(X-X_k)\}$ represents the point mass assigned to Y_i, in which $K_h(u)=h^{-1}K(u/h)$ and h is

a bandwidth. Therefore, under the exponential tilting model, a set of modified EEs for the ith observation is given by

$$\hat{\psi}_M(Y_i, Z_i; \theta) = \delta_i \psi(Y_i, Z_i; \theta) + (1 - \delta_i)\hat{m}_\psi(X_i; \theta, \gamma). \tag{2.7}$$

It can be shown that $n^{-1}\sum_{i=1}^n \hat{\psi}_M(Y_i, Z_i; \theta_0)$ is a set of asymptotically unbiased EEs of θ.

2.2. Maximum empirical likelihood estimator

We assume that the value of γ is known. Although γ may be unknown in practice, we may either fix γ at a prefixed value or calculate a consistent estimator of γ, denoted by $\hat{\gamma}$. For instance, $\hat{\gamma}$ can be computed from an independent survey or a validation sample that is a subsample of the nonrespondents. Then, we can substitute $\hat{\gamma}$ into (2.7) to get $\hat{\psi}_T(Y_i, Z_i; \theta)$. Therefore, we temporarily assume that γ is known.

Let p_i be the probability weight allocated to $\hat{\psi}_M(Y_i, Z_i; \theta)$. The empirical likelihood (Owen (1990)) for θ based on $\hat{\psi}_M(Y_i, Z_i; \theta)$ can be taken as

$$\hat{L}_n(\theta) = \sup\Big\{ \prod_{i=1}^n p_i \mid p_i \geq 0, \sum_{i=1}^n p_i = 1, \sum_{i=1}^n p_i \hat{\psi}_M(Y_i, Z_i; \theta) = 0 \Big\}.$$

The optimal value of p_i is $\hat{p}_i = n^{-1}\{1 + \lambda_{n1}^T(\theta)\hat{\psi}_M(Y_i, Z_i; \theta)\}^{-1}$, where $\lambda_{n1}(\theta)$ is the Lagrange multiplier and satisfies $Q_{n1}(\theta, \lambda_{n1}) = n^{-1}\sum_{i=1}^n \hat{\psi}_M(Y_i, Z_i; \theta)/\{1 + \lambda_{n1}^T(\theta)\hat{\psi}_M(Y_i, Z_i; \theta)\} = 0$. Therefore, the log empirical likelihood ratio function (LELRF) for θ is given by

$$\hat{\ell}_M(\theta) = -2\log\Big\{ \prod_{i=1}^n (n\hat{p}_i) \Big\} = 2\sum_{i=1}^n \log\Big\{ 1 + \lambda_{n1}^T(\theta)\hat{\psi}_M(Y_i, Z_i; \theta) \Big\}. \tag{2.8}$$

Maximizing $-\hat{\ell}_M(\theta)$ leads to the maximum EL estimator (MELM) of θ, denoted by $\hat{\theta}_e$. Under some smoothness condition, $\hat{\theta}_e$ can be obtained by simultaneously solving

$$Q_{n1}(\theta, \lambda_{n1}) = 0 \quad \text{and} \quad Q_{n2}(\theta, \lambda_{n1}) = n^{-1}\sum_{i=1}^n \frac{\lambda_{n1}^T(\theta)\partial_\theta \hat{\psi}_M(Y_i, Z_i; \theta)}{1 + \lambda_{n1}^T(\theta)\hat{\psi}_M(Y_i, Z_i; \theta)} = 0,$$

where ∂_θ denotes partial derivative with respect to θ.

Let X be an auxiliary variable. In practice, some auxiliary information on X may be available, for example, the mean of X is zero or the distribution of X is symmetric. With the auxiliary information, we can improve statistical inference on θ. Specifically, we assume that the auxiliary information of X can be characterized as $E\{A(X)\} = 0$, where $A(X) = (A_1(X), \ldots, A_r(X))^T$ is a known $r \geq 1$ vector (or scalar) function.

To incorporate the auxiliary information on X, the LELRF for θ is defined as

$$\ell_{AU}(\theta) = -2\max\Big\{\sum_{i=1}^{n}\log(np_i)|p_i \geq 0, \sum_{i=1}^{n}p_i = 1, \sum_{i=1}^{n}p_i\hat{\psi}_M(Y_i, Z_i; \theta) = 0, \sum_{i=1}^{n}p_i A(X_i) = 0\Big\}.$$

Let $\Lambda_i(\theta) = (\hat{\psi}_M^T(Y_i, Z_i; \theta), A^T(X_i))^T$. The estimated LELRF for θ based on $\Lambda_i(\theta)$ can be expressed as

$$\hat{\ell}_{AU}(\theta) = 2\sum_{i=1}^{n}\log\{1 + \lambda_{n2}^T(\theta)\Lambda_i(\theta)\}, \tag{2.9}$$

where $\lambda_{n2}(\theta)$ is a $(q+r)\times 1$ Lagrange multiplier vector that satisfies $n^{-1}\sum_{i=1}^{n}\Lambda_i(\theta)/\{1 + \lambda_{n2}^T(\theta)\Lambda_i(\theta)\} = 0$. Maximizing $-\hat{\ell}_{AU}(\theta)$ leads to the MELE of θ, denoted by $\hat{\theta}_{ae}$. Therefore, under some smoothness condition, $\hat{\theta}_{ae}$ can be calculated by simultaneously solving

$$\begin{aligned}
M_{n1}(\theta, \lambda_{n2}) &= n^{-1}\sum_{i=1}^{n}\frac{\hat{\psi}_M(Y_i, Z_i; \theta)}{1 + \lambda_{n21}^T(\theta)\hat{\psi}_M(Y_i, Z_i; \theta) + \lambda_{n22}^T(\theta)A(X_i)} = 0,\\
M_{n2}(\theta, \lambda_{n2}) &= n^{-1}\sum_{i=1}^{n}\frac{A(X_i)}{1 + \lambda_{n21}^T(\theta)\hat{\psi}_M(Y_i, Z_i; \theta) + \lambda_{n22}^T(\theta)A(X_i)} = 0,\\
M_{n3}(\theta, \lambda_{n2}) &= n^{-1}\sum_{i=1}^{n}\frac{\lambda_{n21}^T(\theta)\partial_\theta\hat{\psi}_M(Y_i, Z_i, \theta)}{1 + \lambda_{n21}^T(\theta)\hat{\psi}_M(Y_i, Z_i; \theta) + \lambda_{n22}^T(\theta)A(X_i)} = 0,
\end{aligned}$$

where $\lambda_{n2} = (\lambda_{n21}^T, \lambda_{n22}^T)^T$.

3. Theoretical Results

3.1. Asymptotic properties of MELE for known γ

We first establish the asymptotic properties of MELE and LELRF for known γ. Then, we approximate the asymptotic covariance of MELE. The detailed assumptions and proofs of our results can be found in the Appendix and supplementary materials, respectively. We need some notation. Let $\xrightarrow{\mathcal{L}}$ denote convergence in distribution, and $a^{\otimes 2} = aa^T$ for any vector a. We define several matrices as follows:

$$\begin{aligned}
V_1 &= E\left[\pi(X, Y)^{-1}\{\psi(Y, Z; \theta_0) - m_\psi^0(X; \theta_0)\}^{\otimes 2}\right] + E\left\{m_\psi^0(X; \theta_0)^{\otimes 2}\right\},\\
V_2 &= E\left\{[\delta_i\{\psi(Y_i, Z_i; \theta) - m_\psi^0(X_i; \theta)\} + m_\psi^0(X_i; \theta)]^{\otimes 2}\right\},\\
\Gamma &= E\{\partial_\theta\psi(Y, Z; \theta)\}.
\end{aligned} \tag{3.1}$$

Theorem 1. *Suppose the conditions given in the Appendix hold. Then*

$$\sqrt{n}(\hat{\theta}_e - \theta_0) \xrightarrow{\mathcal{L}} N(0, \Sigma_e) = N(0, \Sigma_1 \Gamma^T V_2^{-1} V_1 V_2^{-1} \Gamma \Sigma_1), \tag{3.2}$$

where $\Sigma_1 = (\Gamma^T V_2^{-1} \Gamma)^{-1}$.

Theorem 1 gives the asymptotic normality of $\hat{\theta}_e$ for the kernel-assisted EE imputation scheme. From (2.1), we have $\pi(X_i, Y_i) = \{1+\exp(-g(X_i))\exp(\gamma Y_i)\}^{-1}$, with $\gamma = -\phi$. On the other hand, $E\{\delta \exp(\gamma Y)|X\} = \exp(g(X))E(1-\delta|X)$. Then, by the kernel regression method and under (2.1) with known parameter $\gamma = -\phi$, a non-parametric estimator of $\pi(X_i, Y_i)$ can be obtained as $\hat{\pi}(X_i, Y_i) = \hat{\pi}_i(\gamma)$, where

$$\hat{\pi}_i(\gamma) = \{1 + \hat{\alpha}(X_i; \gamma)\exp(\gamma Y_i)\}^{-1}, \tag{3.3}$$

with

$$\hat{\alpha}(X_i; \gamma) = \frac{\sum_{j=1}^n (1-\delta_j) K_h(X_j - X_i)}{\sum_{j=1}^n \delta_j \exp(\gamma Y_j) K_h(X_j - X_i)}.$$

Let $\hat{\eta}_i = \delta_i \hat{\pi}_i(\gamma)^{-1}\{\psi(Y_i, Z_i; \theta_0) - \hat{m}_\psi(X_i; \theta_0)\} + \hat{m}_\psi(X_i; \theta_0)$, where $\hat{m}_\psi(X_i; \theta_0) = \hat{m}_\psi(X_i; \theta_0, \gamma)$. Then, a consistent estimator of V_1 is $\hat{V}_1 = n^{-1}\sum_{i=1}^n \hat{\eta}_i^{\otimes 2} - (n^{-1}\sum_{i=1}^n \hat{\eta}_i)^{\otimes 2}$. Furthermore, the consistent estimators of Γ and V_2 are $\hat{\Gamma} = n^{-1}\sum_{i=1}^n \partial_\theta \hat{\psi}_M(Y_i, Z_i; \theta_0)$ and $\hat{V}_2 = n^{-1}\sum_{i=1}^n \hat{\psi}_M(Y_i, Z_i; \theta_0)^{\otimes 2}$, respectively. Thus, Σ_e can be consistently estimated by $\hat{\Sigma}_e = \hat{\Sigma}_1 \hat{\Gamma}^T \hat{V}_2^{-1} \hat{V}_1 \hat{V}_2^{-1} \hat{\Gamma} \hat{\Sigma}_1$, where $\hat{\Sigma}_1 = (\hat{\Gamma}^T \hat{V}_2^{-1} \hat{\Gamma})^{-1}$.

Under the MAR assumption, $\pi(X_i, Y_i)$ reduces to $P(X_i) = \exp\{g(X_i)\}/[1 + \exp\{g(X_i)\}]$. Since $E\{\psi(Y, Z; \theta) - m_{1\psi}(X; \theta)|X\} = 0$, V_1, V_2, and Γ, respectively, reduce to

$$\begin{aligned} V_1 &= E\left\{P(X)^{-1}\Sigma_\psi(X)\right\} + E\left\{m_\psi^0(X; \theta_0)^{\otimes 2}\right\}, \\ V_2 &= E\left\{P(X)\Sigma_\psi(X)\right\} + E\left\{m_\psi^0(X; \theta_0)^{\otimes 2}\right\} \quad \text{and} \quad \Gamma = E\{\partial_\theta m_{1\psi}(X; \theta_0)\}, \end{aligned} \tag{3.4}$$

where $\Sigma_\psi(X) = \mathrm{cov}\{\psi(Y, Z; \theta_0)|X\}$. Thus, Theorem 1 reduces to Theorem 2 of Zhou, Wan, and Wang (2008) under the MAR assumption. When $\pi(X, Y) = 1$, it can be shown that $V_1 = V_2 = E\left\{\psi(Y, Z; \theta)^{\otimes 2}\right\}$, which leads to $\Sigma_e = (\Gamma^T V_2^{-1} \Gamma)^{-1}$ with $\Gamma = E\{\partial_\theta \psi(Y, Z; \theta)\}$, the asymptotic variance of MELE based on the full observations (Qin and Lawless (1994)). Therefore, when $\pi(X, Y)$ is close to 1, the efficiency of MELE based on our proposed kernel-assisted EE imputation scheme is close to that based on the full observations.

Theorem 2. *Suppose the conditions given in the Appendix hold. As* $n \to \infty$*, we have*

$$\hat{\ell}_M(\theta_0) \xrightarrow{\mathcal{L}} \varrho_1 \chi_1^2 + \varrho_2 \chi_2^2 + \cdots + \varrho_q \chi_q^2,$$

where $\chi_k^2 s$ *are independent* χ^2 *variables with one degree of freedom, and the weights* ϱ_i *are the eigenvalues of* $V_2^{-1} V_1$.

Theorem 2 says the asymptotic distribution of $\hat{\ell}_M(\theta_0)$ as a complicated weighted sum of chi-squares. We can use the asymptotic result in Theorem 2 to construct the confidence region of θ. Specifically, let c_α be the $1-\alpha$ quantile of $\varrho_1\chi_1^2+\varrho_2\chi_2^2+\cdots+\varrho_q\chi_q^2$ for $0<\alpha<1$. An approximate $100(1-\alpha)\%$ empirical-likelihood-based confidence region for θ is given by $CI_\alpha^M(\theta)=\{\theta:\hat{\ell}_M(\theta)\le c_\alpha\}$.

To obtain a simple asymptotic distribution, we define an adjusted LELRF as $\hat{\ell}_M^I(\theta_0)=\hat{R}\hat{\ell}_M(\theta_0)$, where $\hat{R}$ is a consistent estimator of $R=q/\mathrm{tr}\{V_2^{-1}V_1\}$ that measures information loss due to the presence of missing data, Zhou, Wan, and Wang (2008). By replacing θ_0 by $\hat{\theta}_e$ in V_1 and V_2, we can get consistent estimators of V_1 and V_2, denoted by $\hat{V}_1$ and $\hat{V}_2$, respectively. When no data are missing, $r(\theta_0)=1$. Moreover, even though $R\sum_{k=1}^q\varrho_k\chi_k^2$ can be well approximated by a $\chi^2(q)$ distribution, the accuracy of such approximation, $\hat{\ell}_M^I(\theta_0)$, also depends on the values of the ϱ_i's.

We develop another adjusted LELRF, denoted by $\hat{\ell}_M^A(\theta_0)$, whose asymptotic distribution is exactly a χ_q^2 distribution,

$$\hat{\ell}_M^A(\theta_0)=\frac{\hat{W}_1}{\hat{W}_2}\hat{\ell}_M(\theta_0), \tag{3.5}$$

where $\hat{W}_1=\mathrm{tr}\{\hat{V}_1^{-1}\hat{\Sigma}\}$ and $\hat{W}_2=\mathrm{tr}\{\hat{V}_2^{-1}\hat{\Sigma}\}$, in which $\hat{\Sigma}=\{\sum_{i=1}^n\hat{\psi}_M(Y_i,Z_i;\theta_0)\}^{\otimes 2}$. Since $\hat{\ell}_M(\theta_0)=\hat{W}_2+o_p(1)$, $\hat{\ell}_M^A(\theta_0)=\hat{W}_1+o_p(1)$ and $\hat{W}_1\xrightarrow{\mathcal{L}}\chi^2(q)$.

With the auxiliary information on X, we use $\hat{\theta}_{ae}$ and $\hat{\ell}_{AU}$ to denote the MELE of θ and the LELRF based on known γ.

Theorem 3. *Suppose the conditions given in the Appendix hold. Then, we have*

(i) $\sqrt{n}(\hat{\theta}_{ae}-\theta_0)\xrightarrow{\mathcal{L}}N(0,\Sigma_{ae})=N(0,\mathcal{C}^{-1}\mathcal{B}^T\mathcal{A}^{-1}V_{1,\mathrm{AU}}\mathcal{A}^{-1}\mathcal{B}\mathcal{C}^{-1})$, *where* $\mathcal{B}=(\Gamma^T,0^T)^T$, $\mathcal{C}=-\mathcal{B}^T\mathcal{A}^{-1}\mathcal{B}$, $D_2=E\left\{A(X)^{\otimes 2}\right\}$, $D_1=E\left\{m_\psi^0(X;\theta_0)A^T(X)\right\}$,

$$\mathcal{A}=\begin{pmatrix}-V_2 & -D_1\\ -D_1^T & -D_2\end{pmatrix}\quad\textit{and}\quad V_{1,\mathrm{AU}}=\begin{pmatrix}V_1 & D_1\\ D_1^T & D_2\end{pmatrix},$$

(ii) $\hat{\ell}_{\mathrm{AU}}(\theta_0)\xrightarrow{\mathcal{L}}\varrho_1^u\chi_1^2+\cdots+\varrho_{r+q}^u\chi_{r+q}^2$, *where the weights* ϱ_k^u *are the eigenvalues of matrix* $V_{2,\mathrm{AU}}^{-1}V_{1,\mathrm{AU}}$, *and* $V_{2,\mathrm{AU}}=-\mathcal{A}$.

Theorem 3 (i) gives the asymptotic normality of $\hat{\theta}_{ae}$ when auxiliary information is available. To estimate Σ_{ae}, we only need to approximate D_1 and D_2, see the consistent estimators of V_1 and V_2 given below Theorem 1. Specifically, we estimate D_1 and D_2 as $\hat{D}_1=n^{-1}\sum_{i=1}^n\hat{\psi}_M(Y_i,Z_i;\theta_0)A^T(X_i)$ and $\hat{D}_2=n^{-1}\sum_{i=1}^n A(X_i)^{\otimes 2}$. It can be shown that $\Sigma_e-\Sigma_{ae}$ is non-negative definite, which indicates that $\hat{\theta}_{ae}$ is asymptotically more efficient than $\hat{\theta}_e$. Moreover, when

auxiliary information on X is available, the amount of information reduction of $\hat{\theta}_{ae}$ compared to that of $\hat{\theta}_e$ does not depend on $\pi(X,Y)$. This result is consistent with that under a simpler setting in Wang and Rao (2002). Theorem 3 (ii) gives the asymptotic distribution of $\hat{\ell}_{\mathrm{AU}}(\theta_0)$ as a weighted sum of chi-squares; we can propose several adjusted LELRFs based on $\hat{\ell}_{\mathrm{AU}}(\theta_0)$ and construct the confidence region of θ. We omit them for the sake of space.

3.2. Asymptotic properties for estimated γ

In many cases, γ is unknown and has to be estimated. We consider that an estimator for γ is computed from an independent survey, or that an estimate is obtained from a validation sample, a subsample of the nonrespondents.

In either case, the resulting semi-parametric modified EEs for the ith observation of θ is

$$\hat{\psi}_T(Y_i,Z_i;\theta)=\delta_i\psi(Y_i,Z_i;\theta)+(1-\delta_i)\hat{m}_\psi(X_i;\theta,\hat{\gamma}). \tag{3.6}$$

where $\hat{m}_\psi(X;\theta,\gamma)$ is defined in (2.6).

It can be shown that $n^{-1}\sum_{i=1}^n\hat{\psi}_T(Y_i,Z_i;\theta_0)$ is a set of asymptotically unbiased EEs of θ. So, we can define the LELRF for θ based on the semi-parametric modified EEs (3.6). We use $\hat{\theta}_T$ and $\hat{\ell}_T$ to denote the MELE of θ and LELRF based on $\hat{\gamma}$, respectively. Assume that $E\{A(X)\}=0$, where $A(X)=(A_1(X),\ldots,A_r(X))^T$ is a known $r\times1$ vector (or scalar) function and let $\tilde{\Lambda}_i(\theta)=(\hat{\psi}_T^T(Y_i,Z_i;\theta),A^T(X_i))^T$. With the auxiliary information on X, we use $\hat{\theta}_{AT}$ and $\hat{\ell}_{AT}$ to denote the MELE of θ and LELRF based on $\hat{\gamma}$.

We first consider that $\hat{\gamma}$ is estimated from an independent survey.

Theorem 4. *Suppose* (C1)–(C8) *hold,* $\sqrt{n}(\hat{\gamma}-\gamma)\xrightarrow{\mathcal{L}}N(0,V_\gamma)$, *and* $\hat{\gamma}$ *is independent of* $\hat{\psi}_M(Y_i,Z_i;\theta)$. *Then*

(i) $\sqrt{n}(\hat{\theta}_T-\theta_0)\xrightarrow{\mathcal{L}}N(0,\Sigma_T)=N(0,\Sigma_1\Gamma^TV_2^{-1}\tilde{V}_1V_2^{-1}\Gamma\Sigma_1)$, *where* $\tilde{V}_1=V_1+H^TV_\gamma H$, $H=E[(1-\delta)(Y-m_0(X))\{\psi(Y,Z;\theta_0)-m_\psi^0(X;\theta_0)\}^T]$ *and* $m_0(X)=E(Y|X,\delta=0)$;

(ii) $\hat{\ell}_T(\theta_0)\xrightarrow{\mathcal{L}}\varrho_1^\gamma\chi_1^2+\varrho_2^\gamma\chi_2^2+\cdots+\varrho_q^\gamma\chi_q^2$, *where the weights* $\varrho_i^\gamma s$ *are the eigenvalues of* $V_2^{-1}\tilde{V}_1$.

Theorem 4 (i) generalizes Theorem 3 of Kim and Yu (2011) from mean functionals to GEEs. To approximate the asymptotic variance of $\hat{\theta}_T$, we only need a consistent estimator of $\tilde{V}_1$, $\hat{V}_1^*(\theta_0)=\hat{V}_1+\hat{H}^T\hat{V}_\gamma\hat{H}$, where $\hat{V}_\gamma$ and $\hat{H}$ are, respectively, consistent estimators of V_γ and H, $\hat{V}_1=n^{-1}\sum_{i=1}^n\hat{\eta}_i^{\otimes2}-(n^{-1}\sum_{i=1}^n\hat{\eta}_i)^{\otimes2}$ with $\hat{\eta}_i=\delta_i\hat{\pi}_i(\hat{\gamma})^{-1}\{\psi(Y_i,Z_i;\theta_0)-\hat{m}_\psi(X_i;\theta_0,\hat{\gamma})\}+\hat{m}_\psi(X_i;\theta_0,\hat{\gamma})$. A consistent estimate of H is given by $\hat{H}=n^{-1}\sum_{i=1}^n(1-\delta_i)\hat{\sigma}_0^2(x_i;\hat{\gamma})$, where $\hat{\sigma}_0^2(X;\hat{\gamma})=\mathcal{Q}^{-1}\sum_{j=1}^n\delta_j\exp(\hat{\gamma}Y_j)K_h(X_j-X)\{Y_j-\hat{m}_0(X_j;\hat{\gamma})\}\{\hat{\psi}(Y_j,Z_j;\theta)-\hat{m}_\psi(X_j;\theta)\}$ in

which $\hat{m}_0(X;\hat{\gamma}) = \mathcal{Q}^{-1}\sum_{i=1}^n \delta_i \exp(\hat{\gamma}Y_i)K_h(X_i - X)Y_i$, $\mathcal{Q} = \sum_{i=1}^n \delta_i \exp(\hat{\gamma}Y_i)$ $K_h(X_i - X)$, and $\hat{m}_\psi(X;\theta,\hat{\gamma}) = \sum_{i=1}^n \omega_{i0}(X;\hat{\gamma})\psi(Y_i, Z_i;\theta)$.

Compared with $\hat{\theta}_e$, $\hat{\theta}_T$ has larger asymptotic variance due to estimating γ. The asymptotic variance of $\hat{\theta}_T$ is the same as that of $\hat{\theta}_e$ when $\hat{\gamma}$ is exactly estimated. Moreover, if $\hat{\gamma}$ is exactly estimated, then $V_\gamma = 0$ and $\tilde{V}_1$ is equal to V_1.

Theorem 5. *Under the conditions of Theorem* 4*, we have*

(i) $\sqrt{n}(\hat{\theta}_{AT} - \theta_0) \xrightarrow{\mathcal{L}} N(0, \Sigma_{AT}) = N(0, \mathcal{C}^{-1}\mathcal{B}^T\mathcal{A}^{-1}\tilde{V}_{1,\mathrm{AU}}\mathcal{A}^{-1}\mathcal{B}\mathcal{C}^{-1})$, *where*

$$\tilde{V}_{1,\mathrm{AU}} = \begin{pmatrix} \tilde{V}_1 & D_1 \\ D_1^T & D_2 \end{pmatrix},$$

$\mathcal{A}$, $\mathcal{B}$, *and* $\mathcal{C}$ *as defined in Theorem* 3.

(ii) $\hat{\ell}_{AT}(\theta_0) \xrightarrow{\mathcal{L}} \varrho_1^a\chi_1^2 + \cdots + \varrho_{r+q}^a\chi_{r+q}^2$, *where the weights* ϱ_i^a*s are the eigenvalues of matrix* $V_{2,\mathrm{AU}}^{-1}\tilde{V}_{1,\mathrm{AU}}$, *and* $V_{2,\mathrm{AU}} = -\mathcal{A}$.

It is can be shown that $\Sigma_{AT} \leq \Sigma_T$ indicating that $\hat{\theta}_{AT}$ based on $\tilde{\Lambda}_i(\theta)$ is asymptotically more efficient than $\hat{\theta}_T$. Thus, the auxiliary information can be used to improve the efficiency of MELE. Theorem 5 generalizes the existing results in Kim and Yu (2011) and Wang and Rao (2002).

We now consider that a validation sample is randomly selected from the set of nonrespondents and responses are obtained for all the elements in the validation sample. A consistent estimator $\hat{\gamma}$ of γ can be obtained by solving

$$\sum_{i=1}^n (1 - \delta_i) r_i \{\psi(Y_i, Z_i;\theta) - \hat{m}_\psi(X_i;\theta,\gamma)\} = 0, \tag{3.7}$$

for γ, where r_i is an indicator of unit i belonging to the follow-up sample, and $\hat{m}_\psi(X_i;\theta,\gamma)$ is defined in (2.6).

Using the estimated titling parameter $\hat{\gamma}$ obtained from (3.7), one can construct $\hat{\psi}_T(Y_i, Z_i;\theta)$ in (3.6). Further, we can construct MELE $\hat{\theta}_T$ and LELRF $\hat{\ell}_T$.

Theorem 6. *Suppose* (C1)−(C8) *hold, except for the semiparametric exponential tilting model in* (2.1)*. Assume that the solution* $\hat{\gamma}$ *to* (3.7) *exists almost everywhere. Let* $\hat{\ell}_T$ *be the LELRF based on the semi-parametric modified EEs* (3.6) *using* $\hat{\gamma}$ *obtained by solving* (3.7) *and the corresponding MELE is* $\hat{\theta}_T$*. Then*

(i)

$$\sqrt{n}(\hat{\theta}_T - \theta_0) \xrightarrow{\mathcal{L}} N(0, \tilde{\Sigma}_T) = N(0, \Sigma_1\Gamma^T V_2^{-1}\tilde{V}_{\mathrm{v}}V_2^{-1}\Gamma\Sigma_1),$$

where $\tilde{V}_{\mathrm{v}} = \mathrm{Var}(\eta_{1i})$,

$$\eta_{1i} = m_\psi^0(X_i;\theta,\gamma_0) + \{\frac{r_i}{\nu}(1-\delta_i)+\delta_i\}\{\psi(Y_i,Z_i;\theta) - m_\psi^0(X_i;\theta,\gamma_0)\},$$

$m_\psi^0(X_i;\theta,\gamma) = \mathrm{pr}\lim_{n\to\infty}\hat{m}_\psi(X_i;\theta,\gamma)$, $\nu = E(r|\delta=0)$ *and* γ_0 *is the probability limit of* $\hat{\gamma}$.

(ii)
$$\hat{\ell}_T(\theta_0) \xrightarrow{\mathcal{L}} \varrho_1^\gamma\chi_1^2 + \varrho_2^\gamma\chi_2^2 + \cdots + \varrho_q^\gamma\chi_q^2,$$

where the weights $\varrho_i^\gamma s$ *are the eigenvalues of matrix* $V_2^{-1}\tilde{V}_{\mathrm{v}}$.

A consistent estimator of $\tilde{V}_{\mathrm{v}}$ is

$$\hat{\tilde{V}}_{\mathrm{v}} = \frac{1}{n}\sum_{i=1}^n \hat{\eta}_{1i}^{\otimes 2} - (\frac{1}{n}\sum_{i=1}^n \hat{\eta}_{1i})^{\otimes 2},$$

with

$$\hat{\eta}_{1i} = \hat{m}_\psi(X_i;\theta,\hat{\gamma}) + \{\frac{r_i}{\nu}(1-\delta_i)+\delta_i\}\{\psi(Y_i,Z_i;\theta) - \hat{m}_\psi(X_i;\theta,\hat{\gamma})\}.$$

In Theorem 6, the exponential tilting model (2.1) is not needed to show (i). The variance $\tilde{V}_{\mathrm{v}}$ can be written as

$$\tilde{V}_{\mathrm{v}} = \mathrm{Var}(\psi(Y,Z;\theta)) + (\nu^{-1}-1)E[(1-\delta)\{\psi(Y,Z;\theta) - m_\psi^0(X;\theta,\gamma_0))\}^{\otimes 2}].$$

Note that

$$m_\psi^0(X;\theta,\gamma) = \mathrm{pr}\lim_{n\to\infty}\hat{m}_\psi(X;\theta,\gamma) = \frac{E\{\delta\psi(Y,Z;\theta)\exp(\gamma Y)|X\}}{E\{\delta\exp(\gamma Y)|X\}},$$

Thus, if (2.1) is true, then $\gamma_0 = \gamma$, and by (i),

$$\begin{aligned} m_\psi^0(X;\theta,\gamma_0) &= \frac{E\{\delta\psi(Y,Z;\theta)\exp(\gamma Y)|X\}}{E\{\delta\exp(\gamma Y)|X\}} = \frac{E\{(1-\delta)\psi(Y,Z;\theta)|X\}}{E\{(1-\delta)|X\}} \\ &= E\{\psi(Y,Z;\theta)|X,\delta=0\} = m_\psi^0(X;\theta). \end{aligned}$$

Since

$$E[(1-\delta)\{\psi(Y,Z;\theta) - m_\psi^0(X;\theta,\gamma_0))\}^{\otimes 2}] \geq E[(1-\delta)\{\psi(Y,Z;\theta) - m_\psi^0(X;\theta))\}^{\otimes 2}],$$

the variance $\tilde{\Sigma}_T$ in (i) is minimized when (2.1) is true. Thus, the validity of the proposed estimator does not depend on the assumed response model and the role of (2.1) is to improve efficiency.

With the auxiliary information on X, we also use $\hat{\theta}_{AT}$ and $\hat{\ell}_{AT}$ to denote the MELE of θ and LELRF based on $\hat{\gamma}$, estimated by the validation sample.

Theorem 7. *Under the conditions of Theorem* 6*, we have*

(i) $\sqrt{n}(\hat{\theta}_{AT} - \theta_0) \xrightarrow{\mathcal{L}} N(0, \tilde{\Sigma}_{AT}) = N(0, \mathcal{C}^{-1}\mathcal{B}^T\mathcal{A}^{-1}\tilde{V}_{\mathrm{v,AU}}\mathcal{A}^{-1}\mathcal{B}\mathcal{C}^{-1})$, *where*

$$\tilde{V}_{\mathrm{v,AU}} = \begin{pmatrix} \tilde{V}_{\mathrm{v}} & D_1 \\ D_1^T & D_2 \end{pmatrix},$$

with $\mathcal{A}$, $\mathcal{B}$ *and* $\mathcal{C}$ *as defined in Theorem* 3 *and* $\tilde{V}_{\mathrm{v}}$ *as defined in Theorem* 6.

(ii) $\hat{\ell}_{AT}(\theta_0) \xrightarrow{\mathcal{L}} \varrho_1^a\chi_1^2 + \cdots + \varrho_{r+q}^a\chi_{r+q}^2$, *where the weights* ϱ_i^a*s are the eigenvalues of* $V_{2,\mathrm{AU}}^{-1}\tilde{V}_{\mathrm{v,AU}}$.

3.3. Bandwidth selection

Let $F_0(y, z|X = x) = P(Y \le y, Z \le z|X = x, \delta = 0)$ be the conditional distribution of (Z, Y) given $X = x, \delta = 0$. Then, based on the exponential tilting model (2.1), a kernel estimator of $F_0(y, z|X = x)$ based on the sample is

$$\begin{aligned}\hat{F}_0(y, z|X = x) &:= \hat{F}_0(y, z|X = x; \gamma) \\ &= \frac{\sum_{j=1}^n \delta_j \exp(\gamma Y_j) I(Y_j \le y) I(Z_j \le z) K_h(x - X_j)}{\sum_{j=1}^n \delta_j \exp(\gamma Y_j) K_h(x - X_j)}.\end{aligned}$$

Then $m_\psi^0(x; \theta) = E\{\psi(Y, Z; \theta)|X = x, \delta = 0\}$ may be estimated by

$$\hat{\mathcal{R}}(x) = \int \psi(y, z; \theta) d\hat{F}_0(y, z|X = x; \gamma).$$

It is known that in nonparametric or semiparametric inferences, selecting a suitable bandwidth is a critical issue. The classical optimal rate for the bandwidth is $h = n^{-1/5}$, see Sepanski, Knickerbocker, and Carroll (1994). But as Zhou, Wan, and Wang (2008) point out, the optimal rate $h = n^{-1/5}$ is not allowed here since we require $nh^{2m} \to 0$ for the mth kernel. Along the lines of Zhou, Wan, and Wang (2008), we suggest the suitable and simple bandwidth $h = \hat{\sigma}_X n^{-1/3}$, where $\hat{\sigma}_X$ is the standard deviation of observation X.

3.4. Reduced Dimension of X

In practical applications the dimension of variate X is high and it is difficult to get an accurate estimator of $m_\psi^0(X_i; \theta)$ by a kernel-smoothing procedure. Here, we propose a dimension reduction technique such that our method is still effective for high-dimensional data.

Let $\mathcal{S}$ be a continuous function from $\mathcal{R}^{d_x}$ to $\mathcal{R}$, such that $\mathcal{S} = \mathcal{S}(X)$ is univariate and $\mathcal{S}_i = \mathcal{S}(X_i)$. Suppose $E\{\delta_i\psi(Y_i, Z_i; \theta)\exp(\gamma Y_i)|\mathcal{S}_i\}/E\{\delta_i \exp(\gamma Y_i)|\mathcal{S}_i\} =$

$E\{\delta_i\psi(Y_i, Z_i; \theta)\exp(\gamma Y_i)|X_i\}/E\{\delta_i \exp(\gamma Y_i)|X_i\}$. Then, if (2.1) is true,

$$\begin{aligned}
&E\{\delta_i\psi(Y_i, Z_i; \theta) + (1-\delta_i)m_\psi^0(\mathcal{S}_i; \theta)\}\\
&= E\Bigg\{\mathrm{pr}(\delta_i = 1|X_i)E(\psi(Y_i, Z_i; \theta)|\delta_i = 1, X_i)\\
&\quad +\mathrm{pr}(\delta_i = 0|X_i)\frac{E\{\delta_i\psi(Y_i, Z_i; \theta)\exp(\gamma Y_i)|\mathcal{S}(X_i)\}}{E\{\delta_i \exp(\gamma Y_i)|\mathcal{S}(X_i)\}}\Bigg\}\\
&= E\Bigg\{\mathrm{pr}(\delta_i = 1|X_i)E(\psi(Y_i, Z_i; \theta)|\delta_i = 1, X_i)\\
&\quad +\mathrm{pr}(\delta_i = 0|X_i)E(\psi(Y_i, Z_i; \theta)|\delta_i = 0, X_i)\Bigg\}\\
&= E\Bigg\{E(\delta_i\psi(Y_i, Z_i; \theta)|X_i) + \mathrm{pr}(\delta_i = 0|X_i)\frac{E\{(1-\delta_i)\psi(Y_i, Z_i; \theta)|\mathcal{S}(X_i)\}}{E\{(1-\delta_i)|X_i\}}\Bigg\}\\
&= E\{\psi(Y_i, Z_i; \theta)\} = 0,
\end{aligned}$$

here $m_\psi^0(\mathcal{S}_i; \theta) = E\{\delta_i\psi(Y_i, Z_i; \theta)\exp(\gamma Y_i)|\mathcal{S}_i\}/E\{\delta_i \exp(\gamma Y_i)|\mathcal{S}_i\}$. Therefore, the resulting EEs can be modified as

$$\hat{\psi}_M(Y_i, Z_i; \theta) = \delta_i\psi(Y_i, Z_i; \theta) + (1-\delta_i)\hat{m}_\psi(\mathcal{S}_i; \theta, \gamma), \tag{3.8}$$

where $\hat{m}_\psi(\mathcal{S}_i; \theta, \gamma)$ is obtained as was $\hat{m}_\psi(X_i; \theta, \gamma)$ in (2.6), except that X is replaced by $\mathcal{S}$. This allows us to deal with the curse-of-dimensionality problem.

4. Numerical Examples

4.1. Simulation studies

Simulation studies of a nonlinear regression model and a logistic regression model were conducted to evaluate the finite sample performance of our proposed MELEs and LELRFs.

Experiment 1. We simulated $\{(X_i, Y_i) : i = 1, \ldots, n\}$ from a nonlinear regression model. Each dataset contained n observations. For each i, X_i was generated from a uniform distribution $U(0, 1)$ and then, given X_i, Y_i was generated from the normal distribution $N(\theta X_i + \exp(\theta X_i), 1)$ with $\theta = 1$. We assumed the X_i's completely observed, but the Y_i's subject to missingness. We generated δ_i, the missing indicator for Y_i, from a Bernoulli distribution with probability $\pi(X_i, Y_i) = P(\delta_i = 1|X_i, Y_i)$. We examined seven missing data mechanisms:

(i) $\pi(X, Y) = 1$ for all X and Y;

(ii) $\mathrm{logit}\{\pi(X, Y)\} = \alpha_0 + \alpha_1 X$ with $(\alpha_0, \alpha_1) = (1.8, 0.5)$;

(iii) $\mathrm{logit}\{\pi(X, Y)\} = \alpha_0 + \alpha_1 X + \alpha_2 Y$ with $(\alpha_0, \alpha_1, \alpha_2) = (1.8, 0.25, 0.15)$;

(iv) $\text{logit}\{\pi(X,Y)\} = \alpha_0 + \alpha_1 X^2 + \alpha_2 Y$ with $(\alpha_0, \alpha_1, \alpha_2) = (1.5, 0.25, 0.5)$;

(v) $\text{logit}\{\pi(X,Y)\} = \alpha_0 + \alpha_1 X + \alpha_2 Y^2$ with $(\alpha_0, \alpha_1, \alpha_2) = (1.5, 0.5, 0.25)$;

(vi) $\text{logit}\{\pi(X,Y)\} = \alpha_0 + \alpha_1 X + \alpha_2 Y + \alpha_3 XY$ with $(\alpha_0, \alpha_1, \alpha_2, \alpha_3)$=$(1.5, 0.15, 0.5, 0.25)$;

(vii) $\text{logit}\{\pi(X,Y)\} = \alpha_0 + \alpha_1 X + \alpha_2 Y + \alpha_3 XY$ with $(\alpha_0, \alpha_1, \alpha_2, \alpha_3) = (0.0001, 0.005, 0.05, 0.25)$.

Scenario (i) is full observation, while (ii) describes a missing at random scenario. Scenarios (iii)−(vii) describe nonignorable missing mechanisms. Scenarios (ii)−(iv) satisfy (2.1) for missing Y. However, (v), (vi), and (vii), which do not satisfy (2.1) and prescribe selection bias in the missingness, were used to investigate the robustness of our proposed empirical likelihood method with respect to the misspecified $\pi(X,Y)$. We took sample size $n = 100$, and simulated 1,000 datasets under each scenario. Then, we created the incomplete data sets for each of 1,000 complete data sets under the six missing data mechanisms. The average missing proportions corresponding to (ii)−(vii) were 11.63%, 9.70%, 7.56%, 6.30%, 6.40%, and 39.45%, respectively.

We considered a set of estimating equations as follows:

$$\psi(Y,X;\theta) = \begin{pmatrix} \psi_1(Y,X;\theta) \\ \psi_2(Y,X;\theta) \end{pmatrix} = \begin{pmatrix} Y^2 - \theta^2/3 - 2\theta X \exp(\theta X) - \exp(2\theta X) - 1 \\ Y - \exp(\theta X) - \theta/2 \end{pmatrix}.$$

For model (2.1), we considered two cases including $\phi = 0$ for the MAR assumption and unknown ϕ. To estimate ϕ in model (2.1), we used a validation sample randomly selected from the set of nonrespondents, Kim and Yu (2011). We chose the Gaussian kernel $K(u) = \exp(-u^2/2)/(2\pi)^{1/2}$ and set the bandwidth h for estimating $m^0_\psi(X;\theta)$ at $c\hat{\sigma}_x n^{-1/3}$, where c is a constant chosen to be 1 in this experiment, and $\hat{\sigma}_x$ is the standard deviation of observations $\{X_i : i = 1, \ldots, n\}$ (Zhou, Wan, and Wang (2008)). We used auxiliary information on X specified by $E(X - 0.5)^2 = 1/12$. We applied the EL method based on the EEs $\psi(Y,X;\theta)$ and model (2.1) to compute the MELEs and 95% confidence intervals of θ. Table 1 presents the results.

Inspecting the results in Table 1 reveals the following. MELEs based on the auxiliary information on X outperformed those without the auxiliary information. When model (2.1) was used and ϕ was estimated, even though the missingness mechanism was misspecified under (ii), (v), (vi), and (vii), the MELEs of $\hat{\theta}$ were close to their true values. Moreover, their empirical coverage levels were relatively close to the pre-specified nominal level 95%. This indicates robustness of the nonignorable missingness model (2.1). Under the MAR assumption, $\phi = 0$ in model (2.1), the MELEs and confidence intervals of θ under (iii)−(vii) were inaccurate. Under (2.1), the confidence intervals for known γ were shorter than

those for estimated γ. As expected, increasing the mean response rates improves the accuracy of parameter estimate and the empirical coverage of confidence interval.

Experiment 2. We simulated $\{(Y_i, X_i) : i = 1, \ldots, n\}$ as follows. We generated $X_{i1} \sim U(0, 2)$ and $X_{i2} \sim N(0, 1)$, and then we simulated $Y_i \sim \text{Bernoulli}(p_i)$, where $\text{logit}\{p_i\} = X_{i1} + 0.5X_{i2}$. We assumed the X_i's completely observed, but the Y_i's subject to missingness. To create missing responses, we generated δ_i for Y_i from a Bernoulli distribution with probability $\pi_i = \pi(X_i, Y_i; \alpha)$ given by $\text{logit}\{\pi(X_i, Y_i; \alpha)\} = \alpha_0 + \alpha_1 X_{i1} + \alpha_2 X_{i2} + \alpha_3 Y_i + \alpha_4 X_{i1} Y_i + \alpha_5 X_{i2} Y_i$, where $\alpha = (\alpha_0, \ldots, \alpha_5)^T$. We considered (i) $\alpha = (1.0, 0.25, 0.20, 0.25, 0.20, 0.20)^T$, and (ii) $\alpha = (1.5, 0.15, 0.20, 0.25, 0.20, 0.20)^T$. Their corresponding average missing proportions were 17.81% and 12.73%, respectively. We took sample size $n = 100$ and simulated 1,000 datasets under each scenario.

We considered the missing mechanism model (2.1) and a set of EEs as follows:

$$\psi(Y_i, X_i; \beta) = (X_{i1}, X_{i2})^T \{Y_i - \text{logit}(\beta_1 X_{i1} + \beta_2 X_{i2})\}, \tag{4.1}$$

where $\beta = (\beta_1, \beta_2)^T$. For (2.1), we considered $\phi = 0$ for the MAR assumption, and unknown ϕ. To estimate ϕ in (2.1), we used a validation sample randomly selected from the set of nonrespondents, Kim and Yu (2011). To estimate $m^0_\psi(X; \theta)$ in which $\theta = \beta$, we took the kernel function to be $K(x_1, x_2) = K(x_1)K(x_2)$, and the bandwidth h to be $\hat{\sigma}_{x_1} n^{-1/3}$, where $K(x) = \exp(-x^2/2)/(2\pi)^{1/2}$ and $\hat{\sigma}_{x_1}$ is the standard deviation of observations $\{X_{i1} : i = 1, \ldots, n\}$. As auxiliary information, we considered $E(X_1 - 1)^2 = 1/3$ and $E(X_2)^2 = 1$. We applied the EL method based on the EEs $\psi(Y, X; \theta)$ and (2.1) to compute the MELEs and 95% confidence intervals of θ. We present the results in Table 2.

Inspecting the results in Table 2 reveals the following. Under (2.1) with estimated γ, the MELEs and confidence intervals of θ were relatively accurate. This indicates that (2.1) is robust to some degree of model misspecification, since the true probability function $\pi(X_i, Y_i; \alpha)$ is different from (2.1). MELEs with the auxiliary information outperformed those without. For the scenarios without the auxiliary information, the empirical coverage probabilities were not close to the pre-specified nominal confidence level 95% when n was small. As expected, increasing the mean response rates increased the accuracy of the empirical coverage probability and decreased the bias and standard deviation (SD) of MELEs and the confidence interval width of θ. MELEs under scenario (i) had smaller root mean square error (RMS) and SD than those under (ii). This indicates that the misspecified missing data mechanism can influence the accuracy of MELE, but such influence is minor. The values of RMS were relatively close to those of SD, indicating that the estimates of the asymptotic variances of MELEs were reasonably accurate even under a misspecified missing data mechanism.

Table 1. Performance of the MELEs and 95% empirical-likelihood-based confidence intervals for θ with $n = 100$ in Experiment 1.

$\pi(X, Y)$	Est.	Bias	RMS	SD	CP	WD
Scenario (i)	$\hat{\theta}_F$	0.023	0.056	0.050	0.896	0.202
	$\hat{\theta}_{FA}$	0.015	0.051	0.049	0.910	0.230
Scenario (ii)	$\hat{\theta}_T$	0.020	0.058	0.054	0.946	0.246
	$\hat{\theta}_{AT}$	0.012	0.053	0.051	0.949	0.245
	$\hat{\theta}_Z$	0.023	0.058	0.053	0.923	0.245
	$\hat{\theta}_{AZ}$	0.015	0.053	0.051	0.909	0.242
Scenario (iii)	$\hat{\theta}_e$	0.023	0.060	0.055	0.947	0.250
	$\hat{\theta}_{ae}$	0.015	0.055	0.053	0.942	0.249
	$\hat{\theta}_T$	0.022	0.057	0.053	0.955	0.252
	$\hat{\theta}_{AT}$	0.014	0.052	0.050	0.948	0.251
	$\hat{\theta}_Z$	0.021	0.069	0.066	0.927	0.243
	$\hat{\theta}_{AZ}$	0.020	0.064	0.060	0.906	0.239
Scenario (iv)	$\hat{\theta}_e$	0.022	0.056	0.051	0.945	0.256
	$\hat{\theta}_{ae}$	0.013	0.051	0.050	0.947	0.254
	$\hat{\theta}_T$	0.022	0.055	0.051	0.955	0.258
	$\hat{\theta}_{AT}$	0.013	0.051	0.049	0.953	0.256
	$\hat{\theta}_Z$	0.016	0.065	0.063	0.934	0.245
	$\hat{\theta}_{AZ}$	0.015	0.060	0.058	0.905	0.242
Scenario (v)	$\hat{\theta}_T$	0.023	0.056	0.051	0.956	0.259
	$\hat{\theta}_{AT}$	0.014	0.051	0.049	0.946	0.257
	$\hat{\theta}_Z$	0.013	0.064	0.063	0.935	0.246
	$\hat{\theta}_{AZ}$	0.012	0.059	0.058	0.919	0.243
Scenario (vi)	$\hat{\theta}_T$	0.022	0.055	0.051	0.959	0.260
	$\hat{\theta}_{AT}$	0.013	0.050	0.049	0.953	0.258
	$\hat{\theta}_Z$	0.017	0.065	0.063	0.933	0.246
	$\hat{\theta}_{AZ}$	0.016	0.060	0.057	0.911	0.243
Scenario (vii)	$\hat{\theta}_T$	0.020	0.094	0.096	0.936	0.407
	$\hat{\theta}_{AT}$	0.023	0.089	0.092	0.935	0.430
	$\hat{\theta}_Z$	0.042	0.063	0.076	0.964	0.375
	$\hat{\theta}_{AZ}$	0.035	0.061	0.070	0.965	0.388

$\hat{\theta}_F$ and $\hat{\theta}_{FA}$ denote MELEs of θ without and with the auxiliary information on X based on the complete data, respectively, whilst $\hat{\theta}_Z$ and $\hat{\theta}_{AZ}$ denote Zhou's estimators in (2.7) with $\gamma = 0$, with and without auxiliary information, 'Bias' denotes the difference between the true value and the mean of the estimates based on 1,000 replications, 'RMS' is the root mean square between the true value and the estimates based on 1,000 replications, 'SD' is the standard deviation of the estimates based on 1,000 replications, 'CP' is the coverage probability and 'WD' is the average interval width.

Table 2. Performance of the MELEs and 95% empirical-likelihood-based confidence intervals for θ in Experiment 2.

$\pi(X,Y)$	β	Est.	Bias	RMS	SD	CP	WD
Full Obs.	β_1	$\hat{\beta}_{1F}$	0.001	0.176	0.176	0.955	0.800
		$\hat{\beta}_{1FA}$	0.001	0.176	0.176	0.962	0.909
	β_2	$\hat{\beta}_{2F}$	0.012	0.195	0.195	0.955	0.800
		$\hat{\beta}_{2FA}$	0.014	0.199	0.198	0.962	0.909
Scenario 1	β_1	$\hat{\beta}_{1T}$	0.019	0.253	0.253	0.941	0.803
		$\hat{\beta}_{1AT}$	0.043	0.211	0.207	0.948	0.782
		$\hat{\beta}_{1Z}$	0.053	0.728	0.727	0.921	0.792
		$\hat{\beta}_{1AZ}$	0.003	0.789	0.789	0.911	0.778
	β_2	$\hat{\beta}_{2T}$	0.021	0.268	0.267	0.941	0.803
		$\hat{\beta}_{2AT}$	0.012	0.251	0.251	0.945	0.782
		$\hat{\beta}_{2Z}$	0.066	0.846	0.844	0.921	0.792
		$\hat{\beta}_{2AZ}$	0.029	0.481	0.480	0.915	0.778
Scenario 2	β_1	$\hat{\beta}_{1T}$	0.032	0.239	0.237	0.951	0.816
		$\hat{\beta}_{1AT}$	0.031	0.196	0.194	0.942	0.791
		$\hat{\beta}_{1Z}$	0.078	0.266	0.254	0.934	0.807
		$\hat{\beta}_{1AZ}$	0.046	0.263	0.259	0.916	0.792
	β_2	$\hat{\beta}_{2T}$	0.032	0.251	0.250	0.951	0.816
		$\hat{\beta}_{2AT}$	0.005	0.234	0.234	0.950	0.791
		$\hat{\beta}_{2Z}$	0.046	0.275	0.272	0.934	0.807
		$\hat{\beta}_{2AZ}$	0.011	0.698	0.699	0.920	0.792

To compare our proposed method with that of Troxel, Lipsitz, and Brennan (1997), we created the missing responses in the 1,000 datasets $\{(Y_i, X_i) : i = 1, \ldots, 100\}$ simulated above. We used the missing data mechanisms

(iii) $\text{logit}\{P(R_{i1} = 1)\} = \alpha_0 + \alpha_1 X_{i1} + \alpha_2 X_{i2} + \alpha_3 Y_i$ and $\text{logit}\{P(R_{i2} = 1|R_{i1} = 0)\} = \alpha_0 + \alpha_1 X_{i1} + \alpha_2 X_{i2} + \alpha_3 Y_i + \tau$, with $(\alpha_0, \alpha_1, \alpha_2, \alpha_3, \tau) = (0.05, 0.25, 0.20, 0.25, 0.5)$,

(iv) $\text{logit}\{P(R_{i1} = 1)\} = \alpha_0 + \alpha_1 X_{i1} + \alpha_2 X_{i2} + \alpha_3 Y_i + \alpha_4 X_{i1} Y_i + \alpha_5 X_{i2} Y_i$ and $\text{logit}\{P(R_{i2} = 1|R_{i1} = 0)\} = \alpha_0 + \alpha_1 X_{i1} + \alpha_2 X_{i2} + \alpha_3 Y_i + \alpha_4 X_{i1} Y_i + \alpha_5 X_{i2} Y_i + \tau$, with $(\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \tau) = (0.5, 0.25, 0.20, 0.25, 0.20, 0.20, 0.5)$.

Here (iii) satisfies (2.1) for missing Y, whilst (iv) does not. The average missing rates corresponding to (iii) and (iv) were, respectively, 38.36% and 26.13%. We computed the estimates of (β_1, β_2) and the 95% confidence regions of (β_1, β_2) by using our proposed EL method and the Troxel, Lipsitz and Brennan (TLB) method based on EEs (4.1) and (2.1).

Table 3 and Figure 1 present the results. Table 3 shows that, compared with the TLB method, our proposed method not only significantly reduced bias, but

Table 3. Estimates of β_1 and β_2 for the EL and TLB methods in Experiment 2.

Scenario (iii)				Scenario (iv)			
Est.	Bias	RMS	SD	Est.	Bias	RMS	SD
$\hat{\beta}_{1T}$	0.074	0.367	0.360	$\hat{\beta}_{1T}$	0.078	0.292	0.282
$\hat{\beta}_{2T}$	0.005	0.306	0.306	$\hat{\beta}_{2T}$	0.006	0.288	0.288
$\hat{\beta}_1^{\mathrm{TLB}}$	0.101	0.305	0.288	$\hat{\beta}_1^{\mathrm{TLB}}$	0.109	0.297	0.276
$\hat{\beta}_2^{\mathrm{TLB}}$	0.026	0.305	0.304	$\hat{\beta}_2^{\mathrm{TLB}}$	0.042	0.308	0.305

also yielded parameter estimates with smaller RMS and SD under (iv), indicating that our proposed method is robust to the misspecified response probability model. Figure 1 shows that our proposed method gave smaller confidence regions than the TLB method.

We suggest that

- EL method can handle over-identified EEs, whereas the TLB method cannot. Moreover, the EL method produces confidence regions, whose shape and orientation are determined entirely by the data. It also does not require a pivotal quantity for constructing confidence regions and has better finite sample performance (Owen (1990)).
- We observed that the TLB method requires correct specification of the missing data mechanisms $(R_{ik}|Y_i, X_i)$ and the model for $(Y_i|X_i)$, which limits its applicability. In contrast, our method does not require a specific form of $F(x, y)$, and the validity of our proposed estimator is robust to the assumed response model $P(R_{i1} = 1|Y_i, X_i)$.

However, our method can suffer from computational difficulties, including optimizing LELRFs and searching the lower and upper limits for confidence regions of parameters. Moreover, our method can break down in the high-dimensional case, which is the topic of our future research.

4.2. A data example

The New York Social Indicators Survey (NYSIS) was a telephone survey of New York City families conducted every two years by the Columbia University School of Social Work. The core survey was designed to document individual and family well-being across multiple domains: human, financial, and social assets; economic and social living conditions; perceptions of the City and its services. The survey also measured the sources and extent of external support from government, family, and friends, community and religious programs, and employers.

A data set was taken from the 2002 NYSIS to illustrate our proposed methodologies. The 2002 SIS survey was conducted between March and June, 2002, and 1501 adults were interviewed. Interviews lasted an average of 24 minutes for

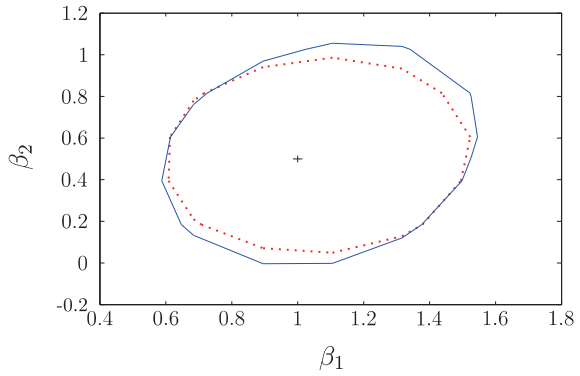

Figure 1. Results of Experiment 2: 95% confidence regions for (β_1, β_2) based on EL (the dot curve) and TLB (the solid curve) under the scenario (iii) with a sample size $n = 100$ for a simulated dataset.

families without children, and 34 minutes for families with children. Let X_{1i} be the number of people in family, X_{2i} be the working hours, and Y_i be the earning of a resident in the New York City in 2001. Since some people were reluctant to report their earnings, some data on Y were missing, but X_{1i} and X_{2i} were fully observed. According to the nature of missing data, we deemed it reasonable to assume the missing data mechanism of Y was non-ignorable. For the 2002 NYSIS data set, the response rate of Y_i's was 89.81%.

Our objective was to use the proposed method to estimate the mean earnings of a resident in the New York City in 2001 and the variance of earnings. The vector of estimating functions is

$$\psi(Y;\theta) = \begin{pmatrix} Y - \theta_1 \\ (Y-\theta_1)^2 - \theta_2^2 \end{pmatrix}.$$

where θ_1 and θ_2^2 are the mean and variance of Y_i's, respectively. Clearly, $E\psi(Y;\theta) = 0$, and we let $\varrho = n^{-1/2}\theta_2$. To obtain the estimator $m_\psi^0(X;\theta)$, we chose $K(x_1,x_2) = K(x_1)K(x_2)$ and set the bandwidth h to be $\hat{\sigma}_x n^{-1/3}$, where $K(x) = \exp(-x^2/2)/(2\pi)^{1/2}$ and $\hat{\sigma}_x$ is the standard deviation of X_1 in the data set. An estimator $\hat{\gamma}$ of the exponential tilting parameter γ was obtained by solving $\sum_{i=1}^n (1-\delta_i) r_i \{\psi(Y_i;\theta) - \hat{m}_\psi(X_i;\theta,\gamma)\} = 0$, where r_i was the indicator of unit i belonging to the follow-up sample, and $\hat{m}_\psi(X;\theta,\gamma)$ is defined in (2.6). The follow-up rate was 25%. To stabilize the computational algorithm, we used 10^{-4} to scale the observed values of the Y_i's. Zhou's estimators, which assume the missing data mechanism is MAR, and our proposed estimators were computed. Results of estimates, standard errors, and 95% confidence intervals for θ_1 and θ_2 are reported in Table 4(a). Table 4(a) has the estimated standard errors (SE) based on Zhou's estimators larger in magnitude than those of our estimators; our

Table 4(a). Estimate (Est), standard error (SE), 95% confidence interval (CI), and average width (AW) for Zhou's estimators (MAR assumption) and our estimators in the case study.

	NMAR			MAR		
Methods	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\varrho}$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\varrho}$
Est	6.204	22.122	0.571	6.132	21.720	0.561
SE	0.560	4.681	—	0.765	6.215	—
NACI	(5.106,7.302)	(12.948,31.296)	—	(4.634,7.631)	(9.539,33.902)	—
NAAW	2.196	18.348	—	2.997	24.363	—
ELCI	(5.354,7.454)	(17.612,26.872)	—	(5.132,7.582)	(17.120,26.590)	—
ELAW	2.100	9.260	—	2.450	9.470	—

Table 4(b). The results when only the variable X_2: "number of working hours" is considered as an auxiliary variable in the case study.

	NMAR			MAR		
Methods	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\varrho}$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\varrho}$
Est	5.711	20.614	0.532	5.765	20.901	0.540
SE	0.529	4.519	—	0.774	6.468	—
NACI	(4.675,6.747)	(11.757,29.470)	—	(4.248,7.282)	(8.223,33.579)	—
NAAW	2.072	17.713	—	3.034	25.356	—
ELCI	(4.911,6.861)	(16.134 25.304)	—	(4.865,7.115)	(16.361,25.711)	—
ELAW	1.950	9.170	—	2.250	9.350	—

NACI denotes the NA-based CI, ELCI represents the EL-based CI, NAAW denotes the average width for the NA-based CI, ELAW represents the average width for the EL-based CI.

proposed estimators had shorter EL-based and NA-based confidence intervals than Zhou's estimators; the EL-based CIs had shorter interval lengths than NA-based CIs; further, results from Table 4(a) indicate that our proposed estimator $\hat{\varrho}$ was, in fact, very close to the estimated standard error of $\hat{\theta}_1$, but there is large bias between $\hat{\varrho}$ and the SE of $\hat{\theta}_1$ based on Zhou's method.

Also, we addressed the case in which only the variable X_2: "number of working hours" is considered as auxiliary. Table 4(b) shows the same conclusions as Table 4(a). Comparing Table 4(b) to Table 4(a), we also find that, based on our proposed method, when only the variable X_2: "number of working hours" is considered as auxiliary, standard errors of the estimates of the θ_1 and θ_2 were smaller than those from considering both X_1 and X_2 as auxiliary and, in this case, the corresponding EL-based and NA-based confidence intervals were shorter. Table 4(b) suggests that we need only consider the variable X_2 as auxiliary to estimate the mean earning of a resident in the New York City in 2001 and the variance of earning in this Indicators Survey.

Acknowledgement

The research was partially supported by grants from the National Science Fund for Distinguished Young Scholars of China (11225103), and Research Fund for the Doctoral Program of Higher Education of China (20115301110004), and NIH grants RR025747-01, P01CA142538-01, MH086633, EB005149-01, and AG033387.

Supplementary material Supplementary Materials available in the attached file include the proofs of Lemmas 1–7 and Theorems 1–7, and the algorithms for computing parameter estimates.

Appendix

Let $f(\cdot)$ be the probability density function of X and $G(X)=f(X)\exp\{g(X)\}\{1-\pi(X)\}$, where $g(X)$ is defined in (2.1). Take $\pi(x,y)=P(\delta=1|X=x,Y=y)$, $\pi(x)=P(\delta=1|X=x)$, $m^0_\psi(x;\theta)=E\{\psi(Y,Z;\theta)|X=x,\delta=0\}$, $m_0(X)=E(Y|X,\delta=0)$, and $m_{Y\psi}(X)=E(Y\psi(Y,Z;\theta)|X,\delta=0)$. The symbol ∂ denotes partial differentiation with respect to parameter θ.

Some regularity conditions are required for the proofs of Theorems 1–7.

(C1) The probability density function $f(x)$ is bounded away from ∞ on the support of X, and the second derivative of $f(x)$ is continuous and bounded.

(C2) The probability function $\pi(X,Y)$ satisfies $\min_i \pi(X_i,Y_i)\geq c_0>0$ a.s. for some positive constant c_0, and $\pi(X)=E(\pi(X,Y)|X)\neq 1$ a.s..

(C3) $E(Y^2)$ and $E\{\exp(2\gamma Y)\}$ are finite.

(C4) $\psi(\cdot;\theta)$ is twice continuously differentiable in the neighborhood of the true value θ_0, and $m_\psi(x;\theta)$ is twice continuously differentiable in the neighborhood of x.

(C5) $0<E|\psi(Y,Z;\theta)|^2<\infty$ and $0<E|\alpha^T\partial_\theta\psi(Y,Z;\theta)|^2<\infty$ for any constant vector α; $\partial_\theta\psi(\cdot;\theta)$ and $\psi^3(\cdot;\theta)$ are bounded by some integrable function $M(z)$ in the neighborhood of θ.

(C6) Matrices V_1, V_2, $\tilde{V}_1$, $\tilde{V}_{\rm v}$, and D_2 are positive definite, and $E\{\partial_\theta\psi(Y,Z;\theta)\}$ has full column rank p.

(C7) The kernel function $K(\cdot)$ is a probability density function such that

(i) it is bounded and has compact support;

(ii) it is symmetric with $\sigma^2=\int\omega^2K(\omega)d\omega<\infty$;

(iii) $K(\omega)\geq d_1$ for some $d_1>0$ in some closed interval centered at zero.

(C8) $nh\to\infty$ and $nh^4\to 0$ as $n\to\infty$.

These assumptions are common in the missing data and nonparametric literatures. Conditions (C2) is similar to that used in Kim and Yu (2011); (C3)−(C6) are standard assumptions for empirical likelihood based inference with estimating equations; (C7) and (C8) are common in the nonparametric literature.

Lemma 1. *Suppose* (C1)−(C8) *hold. Then*

$$n^{-1/2}\sum_{i=1}^{n}\hat{\psi}_M(Y_i,Z_i,\theta)\xrightarrow{\mathcal{L}}N(0,V_1),\quad n^{-1}\sum_{i=1}^{n}\hat{\psi}_M(Y_i,Z_i,\theta_0)^{\otimes 2}\xrightarrow{\mathcal{P}}V_2,$$
$$n^{-1}\sum_{i=1}^{n}\partial_\theta\hat{\psi}_M(Y_i,Z_i,\theta_0)\xrightarrow{\mathcal{P}}\Gamma.$$

Lemma 2. *Suppose* (C1)−(C8) *hold. Then, as* $n\to\infty$*, with probability tending to 1,* $\hat{\ell}_M(\theta)$ *attains its minimum at some point* $\hat{\theta}_e$ *in the interior of the ball* $\| \theta-\theta_0 \|\le n^{-1/3}$*, and the solutions* $\hat{\theta}_e$ *and* $\hat{\lambda}_{n1}=\lambda_{n1}(\hat{\theta}_e)$ *satisfy*

$$Q_{n1}(\hat{\theta}_e,\hat{\lambda}_{n1})=0\quad\textit{and}\quad Q_{n2}(\hat{\theta}_e,\hat{\lambda}_{n1})=0.$$

Lemma 3. *Suppose* (C1)−(C8) *hold. Then*

$$n^{-1/2}\sum_{i=1}^{n}\Lambda_i(\theta_0)\xrightarrow{\mathcal{L}}N(0,V_{1,\mathrm{AU}}),\quad n^{-1}\sum_{i=1}^{n}\Lambda_i(\theta_0)\Lambda_i^T(\theta_0)\xrightarrow{\mathcal{P}}V_{2,\mathrm{AU}},$$

where

$$V_{2,\mathrm{AU}}=\begin{pmatrix}V_2 & D_1\\ D_1^T & D_2\end{pmatrix}.$$

Lemma 4. *Let* U *be* r*-vector of random variables that satisfies* $U\xrightarrow{\mathcal{L}}N(0,I_r)$*, where* I_r *is the* $r\times r$ *identity matrix. Let* P *be a* $r\times r$ *nonnegative definite matrix with eigenvalues* $l_1,\cdots,l_r$*. Then,* $U^TPU\xrightarrow{\mathcal{L}}l_1\chi_1^2+\cdots+l_r\chi_r^2$*, where* χ_i^2*'s* $(i=1,\ldots,r)$ *are* χ^2 *random variables each with one degree of freedom.*

Lemma 5. *Suppose* (C1)−(C8) *hold. Then*

(i) *when the parameter estimate for* γ *is computed from an independent survey,*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{\psi}_T(Y_i,Z_i,\theta)\xrightarrow{\mathcal{L}}N(0,\tilde{V}_1),$$

where $\tilde{V}_1=V_1+H^TV_\gamma H$.

(ii) *when the parameter estimate for* γ *is obtained from a validation sample,*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{\psi}_T(Y_i,Z_i,\theta)\xrightarrow{\mathcal{L}}N(0,\tilde{V}_{\mathrm{v}}),$$

where $\tilde{V}_{\rm v} = {\rm Var}(\eta_{1i})$,

$$\eta_{1i} = m_\psi^0(X_i;\theta,\gamma_0) + \{\frac{r_i}{\nu}(1-\delta_i)+\delta_i\}\{\psi(Y_i,Z_i;\theta) - m_\psi^0(X_i;\theta,\gamma_0)\},$$

$m_\psi^0(X_i;\theta,\gamma) = {\rm pr}\lim_{n\to\infty}\hat{m}_\psi(X_i;\theta,\gamma)$, $\nu = E(r|\delta=0)$, *and* γ_0 *is the probability limit of* $\hat{\gamma}$.

Lemma 6. *Suppose* (C1)−(C8) *hold. Then*

$$\frac{1}{n}\sum_{i=1}^n \hat{\psi}_T(Y_i,Z_i;\theta_0)\hat{\psi}_T^T(Y_i,Z_i;\theta_0) \xrightarrow{\mathcal{P}} V_2, \quad \frac{1}{n}\sum_{i=1}^n \partial_\theta \hat{\psi}_T(Y_i,Z_i;\theta_0) \xrightarrow{\mathcal{P}} \Gamma,$$

where $V_2 = E\left\{[\delta_i\{\psi(Y_i,Z_i;\theta) - m_\psi^0(X_i;\theta)\} + m_\psi^0(X_i;\theta)]^{\otimes 2}\right\}$ *and* $\Gamma = E\{\partial_\theta\psi(Y,Z;\theta)\}$.

Lemma 7. *Suppose* (C1)−(C8) *hold. Then*

(i) *when the parameter estimate for* γ *is computed from an independent survey,*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n \tilde{\Lambda}_i(\theta_0) \xrightarrow{\mathcal{L}} N(0,\tilde{V}_{1,{\rm AU}}), \quad \frac{1}{n}\sum_{i=1}^n \tilde{\Lambda}_i(\theta_0)\tilde{\Lambda}_i^T(\theta_0) \xrightarrow{\mathcal{P}} V_{2,{\rm AU}},$$

where

$$\tilde{V}_{1,{\rm AU}} = \begin{pmatrix} \tilde{V}_1 & D_1 \\ D_1^T & D_2 \end{pmatrix}, \quad V_{2,{\rm AU}} = \begin{pmatrix} V_2 & D_1 \\ D_1^T & D_2 \end{pmatrix}$$

with $D_1 = E\left\{m_\psi^0(X;\theta_0)A^T(X)\right\}$, $D_2 = E\left\{A(X)A^T(X)\right\}$, $\tilde{V}_1$ *is defined in Theorem* 4, *and* V_2 *is defined in Theorem* 2.

(ii) *when the parameter estimate for* γ *is obtained from a validation sample,*

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n \tilde{\Lambda}_i(\theta) \xrightarrow{\mathcal{L}} N(0,\tilde{V}_{{\rm v},{\rm AU}}), \qquad \frac{1}{n}\sum_{i=1}^n \tilde{\Lambda}_i(\theta_0)\tilde{\Lambda}_i^T(\theta_0) \xrightarrow{\mathcal{P}} V_{2,{\rm AU}},$$

where

$$\tilde{V}_{{\rm v},{\rm AU}} = \begin{pmatrix} \tilde{V}_{\rm v} & D_1 \\ D_1^T & D_2 \end{pmatrix}, \quad V_{2,{\rm AU}} = \begin{pmatrix} V_2 & D_1 \\ D_1^T & D_2 \end{pmatrix}$$

$\tilde{V}_{\rm v}$ *is defined in Theorem* 6.

References

Ibrahim, J. G., Chen, M. H., Lipsitz, S. R. and Herring, A. H. (2005). Missing data methods for generalized linear models: a comparative review. *J. Amer. Statist. Assoc.* **100**, 332-346.

Ibrahim, J. G. and Molenberghs, G. (2009). Missing data methods in longitudinal studies: a review. *Test* **18**, 1-43.

Kim, J. K. and Yu, C. L. (2011). A semiparametric estimation of mean functionals with nonignorable missing data. *J. Amer. Statist. Assoc.* **106**, 157-165.

Lee, A. J. and Scott, A. J. (1986). Ultrasound in ante-natal diagnosis. In *The Fascination of Statistics*. (Edited by R. J. Brook, G. C. Arnold, T. H. Hassard, R. M. Pringle), 277-293. Marcel Dekker, New York.

Lipsitz, S. R., Ibrahim, J. G. and Zhao, L. P. (1999). A new weighted estimating equation for missing covariate data with properties similar to maximum likelihood. *J. Amer. Statist. Assoc.* **94**, 1147-1160.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis With Missing Data*. 2nd edition. Wiley.

Newey, W. K. and Smith, R. J. (2004). Higher order properties of gmm and generalized empirical likelihood estimations. *Econometrica* **72**, 219-255.

Owen, A. B. (1990). Empirical likelihood confidence regions. *Ann. Statist.* **18**, 90-120.

Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *Ann. Statist.* **22**, 300-325.

Qin, J., Zhang, B. and Leung, D. (2009). Empirical likelihood in missing data problems. *J. Amer. Statist. Assoc.* **104**, 1492-1503.

Robins, J., Rotnitzky, A. and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.* **89**, 846-866.

Robins, J., Rotnitzky, A. and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J. Amer. Statist. Assoc.* **90**, 106-121.

Seber, G. A. F. and Wild, C. J. (1989). *Nonlinear Regression*. Wiley, New York.

Sepanski, J. H., Knickerbocker, R. and Carroll, R. J. (1994). A Semiparametric Correction for Attenuation. *J. Amer. Statist. Assoc.* **89**, 1366-1373.

Troxel, A. B., Lipsitz, S. R. and Brennan, T. A. (1997). Weighted estimating equations with nonignorably missing response data. *Biometrics* **53**, 857-869.

Wang, D. and Chen, S. X. (2009). Empirical likelihood for estimating equations with missing values. *Ann. Statist.* **37**, 490-517.

Wang, Q. H. and Rao, J. N. K. (2002). Empirical likelihood-based inference under imputation for missing response data. *Ann. Statist.* **30**, 896-924.

Zhou, Y., Wan, A. T. K. and Wang, X. J. (2008). Estimating equations inference with missing data. *J. Amer. Statist. Assoc.* **103**, 1187-1199.

Zhu, H. T., Ibrahim, J. G. Tang, N. S. and Zhang, H. P. (2008). Diagnostic measures for empirical likelihood of general estimating equations. *Biometrika* **95**, 489-507.

Department of Statistics, Yunnan University, Kunming 650091, China.

E-mail: nstang@ynu.edu.cn

Department of Statistics, Yunnan University, Kunming 650091, China.

E-mail: pyzhao@live.cn

Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA.

E-mail: hzhu@bios.unc.edu

(Received August 2012; accepted March 2013)